University of Zurich[UZH]

Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2010

Sound recognition with spiking silicon cochlea and Hidden Markov Models

Jaeckel, D ; Moeckel, R ; Liu, S C

Abstract: In this paper we explore the capabilities of a sound recognition system that combines both a novel bio-inspired custom silicon cochlea chip and a classical Hidden Markov Model (HMM). The cochlea chip front-end produces a form of representation that is analogous to the spike outputs of the biological cochlea. The system is trained with either of 2 target sounds (a clap or a bass drum) in the presence of different levels of white noise or colored noise. We provide experimental results that show 1) the system is able to detect a clap or a bass drum sound even if the amplitude of the target sound was not part of the training set and 2) the performance of the system in detecting a target sound in the presence of white noise or colored noise is around 90% for signal-to-noise ratios down to at least 0.8.

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-47180
Conference or Workshop Item

Originally published at:
Jaeckel, D; Moeckel, R; Liu, S C (2010). Sound recognition with spiking silicon cochlea and Hidden Markov Models. In: Ph.D. Research in Microelectronics and Electronics (PRIME), 2010 Conference, Berlin, 18 July 2010 - 21 July 2010.

Sound Recognition with Spiking Silicon Cochlea and Hidden Markov Models

David Jäckel
Bio Engineering Laboratory,
Dept. of Biosystems Science
and Engineering, ETH Zürich
Email: david.jaeckel@bsse.ethz.ch

Rico Moeckel
Institute of Neuroinformatics,
University of Zürich and ETH Zürich,
Zürich, Switzerland.
Email: moeckel@ini.phys.ethz.ch

Shih-Chii Liu
Institute of Neuroinformatics,
University of Zürich and ETH Zürich,
Zürich, Switzerland.
Email: shih@ini.phys.ethz.ch

***Abstract*—In this paper we explore the capabilities of a sound recognition system that combines both a novel bio-inspired custom silicon cochlea chip and a classical Hidden Markov Model (HMM). The cochlea chip front-end produces a form of representation that is analogous to the spike outputs of the biological cochlea. The system is trained with either of 2 target sounds (a clap or a bass drum) in the presence of different levels of white noise or colored noise. We provide experimental results that show 1) the system is able to detect a clap or a bass drum sound even if the amplitude of the target sound was not part of the training set and 2) the performance of the system in detecting a target sound in the presence of white noise or colored noise is around 90% for signal-to-noise ratios down to at least 0.8.**

I. Introduction

Biological systems are more efficient than present machines in navigating around in natural environments. This is one of the key motivations that have prompted designs of analog Very Large Scale Integrated (aVLSI) sensor chips like silicon cochleas (e.g. [1], [2], [3], [4]) that emulate the structure of their biological counterparts. We are starting to see the appearance of VLSI sensors with spiking outputs that are representative of how sensory outputs are transmitted to upper brain areas of many animals [5], [6]. The processing using this form of signal representation might provide insights into how biological systems can perform better than machines [7].

In this paper we describe a sound recognition system that will be eventually implemented on a robotic platform which has silicon spiking cochleas as the front-end acoustic sensors. The recognition is performed by extracting features from the spike outputs in response to different acoustic sounds and training a Hidden Markov Model (HMM) on these features. We use HMMs for the recognition stage as it has proven useful for a wide range of acoustic tasks [8], [9]. One of the first tasks to which HMMs were applied, was in a sound recognition task [10]. We intend that our sound recognition system will allow the robot to locate and follow certain target sounds even in noisy environments.

In section II we give a short introduction on the biological cochlea so that the reader has an understanding of the biological model that is implemented on the silicon cochlea which is described in section III. In section IV we present the HMMs used in the sound recognition experiments described in section V. Section VI concludes the paper.

II. The Biological Cochlea

The cochlea is a part of the inner ear that plays a central role in hearing [11]. The organ is filled with a fluid that moves in response to the vibrations caused by incoming sound signals to the ear. As the fluid moves it causes the basilar membrane to vibrate. Thousands of hair cells on the membrane sense the vibration in the fluid and excite the spiral ganglion cells which generate so-called action potentials or spikes that travel along nerve fibres to higher-order auditory brain areas.

Because of the physical properties of the basilar membrane, high-frequency inputs activate the basilar membrane closest to the entrance of the cochlea while low-frequency signals travel further down the basilar membrane thus activating inner-hair cells further away from the cochlea's entrance. This spatial arrangement of tone perception is called a tonotopic map [12].

III. Analog VLSI Silicon Cochlea

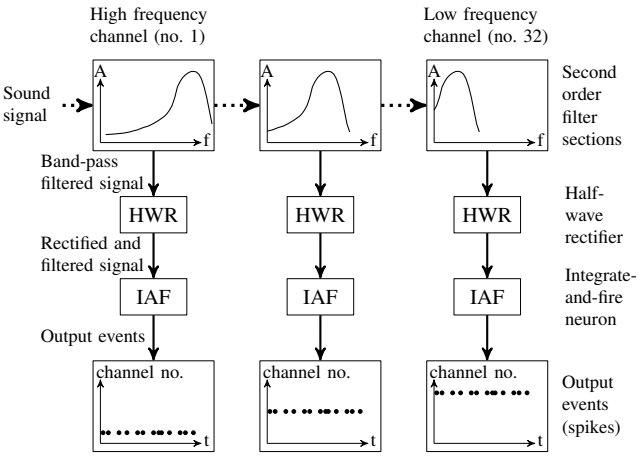

Fig. 1. Schematic of the silicon aVLSI cochlea. The incoming sound signal is processed by a cascade of 32 second-order filter sections; each tuned to a particular best center frequency. The best frequency selectivities of the filter sections are logarithmically distributed from about 100Hz to 2kHz. The band-pass filtered output signals of the individual sections drive a half-wave recifier (HWR) circuit which models the inner hair cells of the cochlea. The resulting half-wave rectified outputs drive integrate-and-fire (IAF) neurons that output events similar to the action potentials of their biological counterparts: the spiral ganglion cells. These output events can be visualized in individual cochleagrams at the bottom of each filter stage.

A schematic of the silicon cochlea used in our experiments is shown in Fig. 1. The detailed circuit description can be found in [6]. The microphone output is processed by a cascade of 32 second-order filter sections in the cochlea. These filter sections model the behaviour of the basilar membrane within the biological cochlea. Each filter section outputs an analog signal corresponding to a band-pass filtered version of the input sound signal. Similar to the biological cochlea, the filter sections are tuned to different characteristic frequencies and arranged to form a tonotopic map. The filter sections that process the input signal first are sensitive to higher frequencies while the sections at the end of the filter cascade are most sensitive to the low frequencies.

The analog output signals of the second-order sections go to individual half-wave rectifier circuits and the subsequent half-wave rectified analog output of each stage drives a leaky integrate-and-fire neuron circuits that models the spiral ganglion cells in the biological cochlea. The output digital events of the neurons are similar to the action potentials generated by the ganglion cells and can be visualized in cochleagrams (for example, Fig. 1, bottom). Each point in the cochleagram refers to an output event that was generated by a certain cochlea channel (ordinate) at a certain point in time (abscissa).

The silicon cochlea thus provides output signals similar to its biological counterpart: It uses a place code where certain neurons or groups of neurons are activated by the auditory input in certain frequency bands. The neuron circuits generate digital output signals where the frequency of the output events reflects the power of the sound signal within the corresponding frequency band: The output event rate increases with the power of the band-pass filtered output.

IV. Sound Recognition Model

A Hidden Markov Model (HMM) is a statistical method that assumes that a system can be modeled with a sequence of hidden states. The term "hidden" refers to the fact that the internal states of the HMM are not visible and cannot necessarily directly be mapped to the visible observations of the system that should be modeled [10].

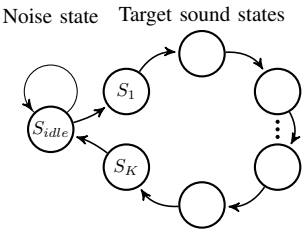

Fig. 2. State diagram of our HMM model. K=60 for the clap sound, K=120 for the bass drum sound.

For our sound recognition experiments we chose two distinct target sounds that can easily be generated: a hand clap which activates most of the cochlea channels in a transient manner and a bass drum sound which activates primarily the lower frequency channels. The corresponding cochleagrams of the 2 target sounds for different amplitudes are shown in Fig. 3.

To model these two target sounds we chose a simplified version of the general HMM (Fig. 2). We model a target sound using a chain of states, S_1 to S_K, and one common idle state S_{idle}. Each state models a certain time step or syllable of the sound. The state S_{idle} is the only state that is connected back to itself and accounts for the fact that a sound sequence can contain several timesteps where no target sound is present. Each state S_i is connected to the next state S_{i+1} with a transition probability of 1. We bin the test sounds into 1 ms timesteps. Since the hand clap and the bass drum sounds have different time lengths (60 ms and 120 ms, respectively), the individual chains also have a different number of states (K=60 and K=120, respectively). The training of a HMM for a particular target sound is done by finding the most likely state transitions and output probabilities given the output sequence, that is, the spike patterns generated by the cochlea. These probabilities can be determined by using an Expectation-Maximization algorithm like the Baum-Welch algorithm [13]. The evaluation of a sound sequence and the detection of a particular target sound is done by finding the most likely state sequence that generated a given output sequence, given the learned probability parameters and the new output sequence consisting of spike patterns from the cochlea. This evaluation is done using the Viterbi algorithm [14], [15].

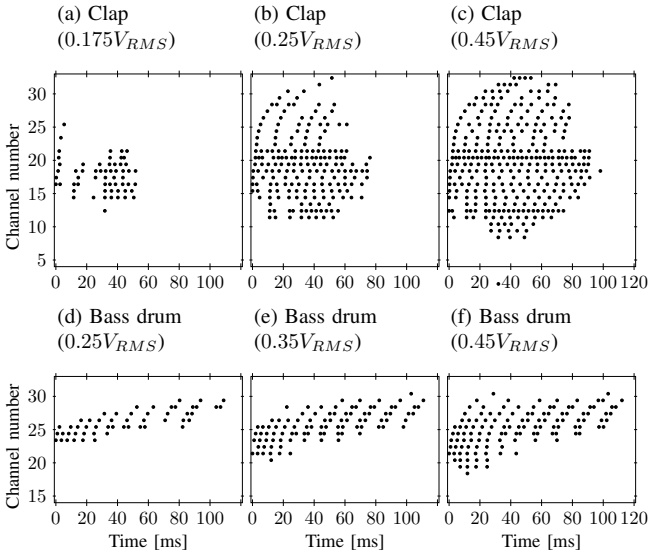

Fig. 3. Cochleagrams comparing clap sounds and bass drum sounds with different input amplitudes in the presence of zero noise (a) Clap ($175mV_{RMS}$). (b) Clap ($250mV_{RMS}$, used during training). (c) Clap ($450mV_{RMS}$). (d) Bass drum ($250mV_{RMS}$). (e) Bass drum ($350mV_{RMS}$, used during training). (f) Bass drum ($450mV_{RMS}$).

V. Results of Sound Recognition Experiments

A. Methods

To evaluate the sound recognition capabilities of the cochlea-HMM combination, we measured the number of correct detection of the target sound (true positives = TP), the

number of target sounds that were detected in the absence of the target sound (false positives = FP), and the number of undetected target sounds (false negatives = FN) for different background noise levels. As a measure of the noise level, we also measured the signal-to-noise ratio (SNR) of the input.

B. Training

We evaluated the capabilities of the cochlea-HMM system in recognizing the 2 target sounds in the presence of both colored noise, that is, noise recorded in a bar, and white noise. The HMM parameters were computed during four **training** sessions.

1) In session 1 we presented the clap sound with an amplitude of $250mV_{rms}$ and added colored noise sound with varying sound levels between $0mV_{rms}$ and $800mV_{rms}$ in steps of $50mV_{rms}$.
2) In session 2 we presented the clap sound with an amplitude of $250mV_{rms}$ and added white noise with varying sound levels between $0mV_{rms}$ and $500mV_{rms}$ in steps of $50mV_{rms}$.
3) In session 3 we presented the bass drum sound with an amplitude of $350mV_{rms}$ and added colored noise with varying sound levels between $0mV_{rms}$ and $800mV_{rms}$ in steps of $50mV_{rms}$.
4) In session 4 we presented the bass drum sound with an amplitude of $350mV_{rms}$ and added white noise with varying sound levels between $0mV_{rms}$ and $500mV_{rms}$ in steps of $50mV_{rms}$.

During the **evaluation** phase in the sound recognition task, we generated new test sequences that were not used during training. Each test sequence lasted 20 seconds and included 12 target sounds of the same type (either claps or bass drum) and of constant RMS amplitude presented at random times in the sequence.

C. Sound recognition performance for untrained sound levels

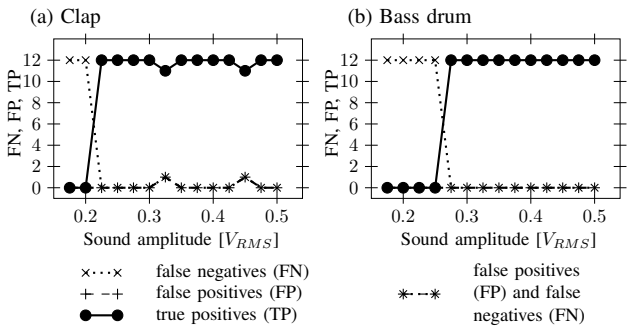

Fig. 4. Test results for untrained sound amplitudes for (a) clap sounds without noise and (b) bass drum sounds without noise. We trained the HMM for clap sounds with an amplitude of $250mV_{RMS}$ and bass drum sounds with an amplitude of $350mV_{RMS}$. During testing, we presented the cochlea-HMM sound recognition system with target sound amplitudes ranging from $250mV_{rms}$ to $500mV_{rms}$.

We first tested the detection performance of the trained cochlea-HMM system on both target sounds with input levels different from that used during training. These experiments were done to evaluate the invariance of the system detection performance to different sound levels. The cochleagrams of the target sounds for 3 different input levels are shown in Fig. 3.

Figures 4(a) and (b) show that the cochlea-HMM system could reliably detect the targets for sound amplitudes up to $500mV_{rms}$. In the case of the clap, the amplitude can be increased up to at least 200% while achieving a detection performance of 90%. In the case of the bass drum, the amplitude can be decreased to 80% of the trained input amplitude and increased to at least 140% while achieving a detection performance of at least 90%.

D. Sound recognition performance in noisy environments

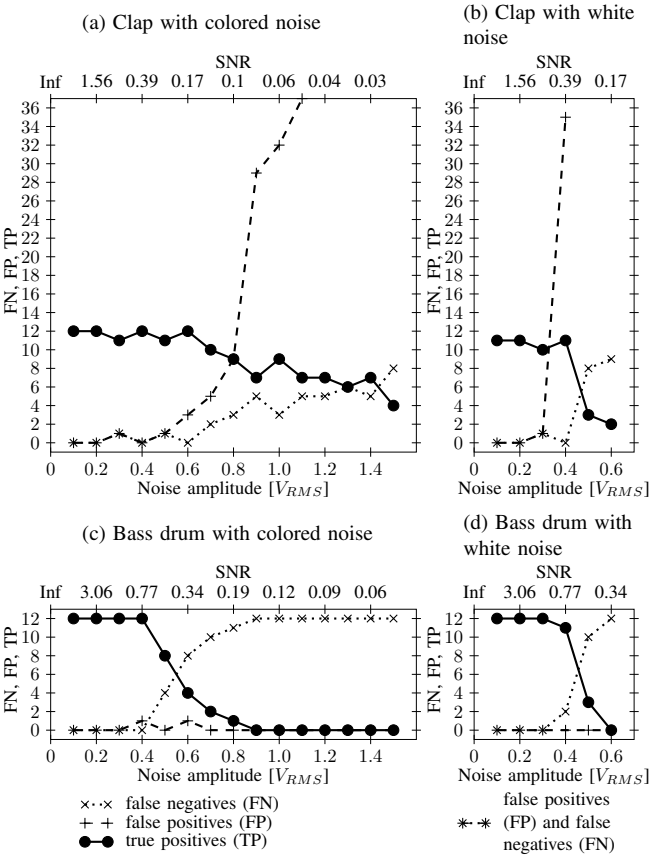

Fig. 5. True positives (TP), false negatives (FN) and false positives (FP) for the HMM recognition of clap and bass drum sounds corrupted by background noise ranging from noise recorded in a busy bar and different levels of white noise. The test sequences contained 12 target sounds.

We used the same amplitudes for the target sounds as during training. The amplitude of the background colored noise was varied between $100mV_{rms}$ and $1500mV_{rms}$ and that of the white noise was varied between $100mV_{rms}$ and $600mV_{rms}$. The maximum test noise amplitudes were larger than the maximum noise amplitudes during training.

The results in Fig. 5(a) show that the clap sound in the presence of colored noise, could be reliably detected even down to a SNR of 0.25. For SNRs < 0.25, fewer target signals

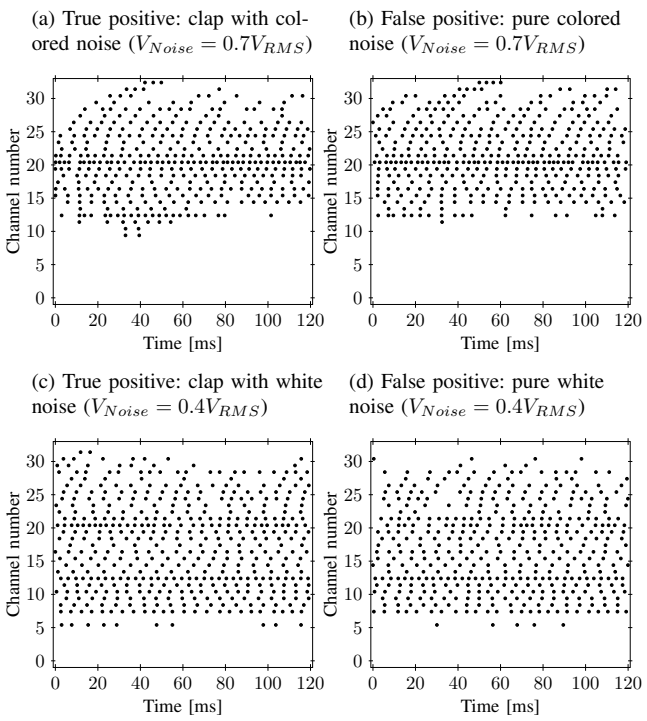

Fig. 6. Comparison of cochleagrams from clap sounds of $250mV_{RMS}$ in the presence of both colored and white noise and from pure noise without a target sound. (a) Correctly detected clap sound (true positive) with colored noise of $700mV_{RMS}$. (b) False positive for pure colored noise of $700mV_{RMS}$. (c) Correctly detected clap sound (true positive) with white noise of $400mV_{RMS}$. (d) False positive for pure white noise of $400mV_{RMS}$.

were detected and the number of false positives increased rapidly. However, a SNR of 0.25 already means that the RMS noise amplitude is two times bigger than that of the target signal. We thus conclude that the cochlea-HMM system was able to reliably recognize the target signal even though the signal power of the noise was four times greater than the power of the target sound.

Figure 5(b) shows the results for the detection of the clap in the presence of white noise. Here the cochlea-HMM system fails for SNR < 0.8. This decrease in performance is probably due to the fact that the white noise with a flat spectrum triggers all cochlea channels, thus leading to the increase in the number of false positives. The cochleagrams in Fig. 6 show that the cochleagram for the clap with colored noise (a) and white noise (c) looks very similar to the cochleagram for pure colored noise (b) and pure white noise (d), thus explaining the reason for the detection performance of the system.

Figures 5(c) and (d) show the detection results for the bass drum sound in the presence of both colored noise and white noise. Interestingly the detection results for types of noise look fairly similar. While the experiments with the clap in the presence of noise showed a strong increase in the detection of false positives as the SNR decreases, almost no false positives were detected in the bass drum experiments. For both types of noise, the bass drum sound could be reliably detected even for SNRs down to 0.77.

VI. Conclusion

We aim at building a sound recognition system based on a novel auditory pre-processor which produces non-framed based data. This detection system uses a silicon cochlea front-end which generate spikes that subsequently drive a Hidden Markov Model (HMM). We show experimental data that characterize the performance of this cochlea-HMM combination in recognizing two target sounds in the presence of white noise and colored noise. We showed two sets of results. In the first case, we showed a sound recognition performance of at least 90% even when the amplitude of the clap sound was varied between 90% to 200% and that of the bass drum was varied between 80% to 140% of the trained amplitude. In the second case, we showed that the system can detect the targets even in the presence of low SNR conditions (SNR$>$0.8).

References

[1] R. F. Lyon and C. A. Mead, "An analog electronic cochlea," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 36, no. 7, pp. 1119–1134, July 1988.

[2] L. Watts, D. A. Kerns, R. F. Lyon, and C. A. Mead, "Improved implementation of the silicon cochlea," *IEEE Journal of Solid-State Circuits*, vol. 27, no. 5, pp. 692–700, May 1992.

[3] A. van Schaik, E. Fragniere, and E. Vittoz, "Improved silicon cochlea using compatible lateral bipolar transistor," in *Advances in Neural Information Processing Systems 11*, D. Touretzky, M. Mozer, and M. Hasselmo, Eds. Cambridge, MA: MIT Press, 1996, pp. 671–677.

[4] S. Mandal, S. M. Zhak, and R. Sarpeshkar, "A bio-inspired active radio-frequency silicon cochlea," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 6, June 2009.

[5] B. Wen and K. Boahen, "A 360-channel speech preprocessor that emulates the cochlear amplifier," in *International Solid-State Circuits Conference Digest of Technical Papers*, February 2006, pp. 556–557.

[6] V. Chan, S.-C. Liu, and A. van Schaik, "AER EAR: A matched silicon cochlea pair with address event representation interface," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 54, no. 1, pp. 48–59, January 2007.

[7] T. Yu, A. Schwartz, J. Harris, M. Slaney, and S.-C. Liu, "Periodicity detection and localization using spike timing from the AER EAR," in *Proceedings of the 2009 IEEE International Symposium on Circuits and Systems*, May 2009, pp. 109–112, ISCAS 2009: Taipei, Taiwan, 24 May–27 May.

[8] S. Yamamoto, J.-M. Valin, K. Nakadai, J. Rouat, F. Michaud, T. Ogata, and H. G. Okuno, "Enhanced robot speech recognition based on microphone array source separation and missing feature theory," in *Proc. International Conference on Robotics and Automation*, 2005.

[9] H. G. Okuno, T. Ogata, K. Komatani, and K. Nakadai, "Computational auditory scene analysis and its application to robot audition," *International Conference on Informatics Research for Development of Knowledge Society Infrastructure*, vol. 0, pp. 73–80, 2004.

[10] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speach recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

[11] A. F. Jahn and J. Santos-Sacchi, *Physiology of the Ear*. San Diego: Singular, 2001.

[12] I. Tasaki, "Nerve impulses in individual auditory nerve fibers of guinea pig," *Journal of Neurophysiology*, vol. 17, no. 2, pp. 97–122, 1954.

[13] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *The Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 167–171, 1970.

[14] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, April 1967.

[15] G. D. Forney, "The Viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, March 1973.